A BENCHMARK DATASET FOR PERFORMANCE EVALUATION OF SHAPE-FROM-X ALGORITHMS

Bellmann Anke, Hellwich Olaf, Rodehorst Volker, Yilmaz Ulas

Berlin University of Technology, Sekr. 3-1, Franklinstr. 28/29, 10587 Berlin, Germay -
(bellmann, hellwich, vr, ulas)@cs.tu-berlin.de

Commission III/2

KEY WORDS: Surface Reconstruction, Shape-From-X, Benchmarking, Dataset, Evaluation

ABSTRACT:

In this study, we present our benchmark dataset for performance evaluation of shape-from-X algorithms and a test procedure for evaluating the reconstruction results. Our aim is to support an objective comparison of different surface reconstruction approaches and to provide an informative basis for the combination of reconstruction methods.

1 INTRODUCTION

The goal of this study is to investigate the performace characteristics of various automatic surface reconstruction algorithms that of class shape-from-X. These algorithms get a collection of calibrated images of an object or a scene and extract 2,5D or 3D shape information as result. Here, *X* denotes the cue used to infer shape. A list of the algorithms that are under investigation in the framework of this study is given in Table 1.

shape-from-X	algorithm
stereo	trinocular-stereo narrow-baseline-stereo wide-baseline-stereo motion (multiview-stereo)
photoconsistency	silhouette (space carving) photoconsistency (voxel coloring) shadow (shadow carving)
texture	texture
shading	shading
focus	(de)focus

Table 1: List of Algorithms

Although many robust implementations of each algorithm are already available in the literature and in practice, there exist hardly any methods to test and compare the performance of these algorithms, quantitatively (Foerstner, 1996). The experimental setup is challenging due to controversial requirements of each algorithm. For instance, the reflectance based methods (Klette et al., 1999) deal with curved Lambertian surfaces, whereas image matching approaches (Courtney et al., 1996, Scharstein and Szeliski, 2002, Seitz et al., 2006) in general prefer textured and/or piecewise planar objects. A comparative study is also difficult due to various reasons. For instance, binocular stereo produces dense depth or disparity maps of the object; whereas multi-view stereo reconstructs the object surface as a 3D polygon mesh or unstructured point cloud and finally reflectance-based methods give surface orientations instead of depth information. In this study, we established a true benchmark dataset for the performance evaluation of shape-from-X algorithms, using a combination of diffuse objects with different surface geometries and synthetic projected textures. We also propose a method for automatically evaluating the results obtained through these algorithms. The same scene is used to acquire input images for each algorithm. So, it enables us to compare 2.5D reconstruction results from one reference view to obtain a measure of success and a ranking. However, the ranking should not be considered as a direct measure of success of one method over another. Instead, it allows an objective comparison of different approaches and provides an informative basis for the combination of reconstruction methods. The limitations of each method are already surveyed and can be found in the literature. However, our aim in establishing this dataset is to provide researchers with a tool, by which they can see how successful or unsuccessful their method with respect to other methods is.

The organization of the paper is as follows. In the following section, the benchmark dataset is explained in detail. An analysis of the surface reconstruction algorithms, which are under consideration, is given in Section 3. The evaluation methodology and ranking strategies are discussed in Section 4. The paper concludes with Section 5 in which we discuss the presented study and state possible improvements.

2 BENCHMARK DATA

Our benchmark dataset consists of

- 360 color images of a real scene,
- 20 synthetic rendered scene images,
- 4 real and 38 synthetic texture patterns,
- orientation and calibration data for
  - 52 camera positions,
  - a texture-mapping LCD-projector,
  - 3 light source directions,
- ground truth in terms of
  - a 3D point cloud,
  - a depth map,
  - a surface orientation map.

2.1 Scene

Finding a single representative object that satisfies the requirements of all reconstruction algorithms is the first challenge in creating the benchmark data. The surface of a representative object should consist of smooth curved parts as well as piecewise

planar patches; on one hand it should have a Lambertian surface, on the other enough textural information on it. In addition to these, the object must be mobile, so that capturing images from various views, is possible. Therefore, we have assembled various objects with different surface geometries, in order to obtain the scene as shown in Figure 1. Object surface is made of white plaster, which gives a good Lambertian surface with perfectly matte reflectance properties. Texture, which is necessary to solve the correspondence problem, is projected using an LCD-projector. This allows analyzing the effect of different texture patterns on the final result, too. Also, artificial shadows are introduced to penalize texture-based approaches. As a result, the final scene is structured enough to rank success of different reconstruction results, as well as general enough to be used by a large variety of algorithms.

Figure 1: The scene

2.2 Calibration

2.2.1 Camera Images are captured using a digital single-lens reflect camera, *Canon EOS-1D Mark II*, which has a 50 mm lens and a 28.7×19.1 mm CMOS sensor. Maximum image size is 3504×2336 pixels. For our experiments this resolution is reduced to 1728×1152, and the captured images are cropped to a region of interest of 1360×904 pixels. Interior orientation of the camera is computed with the bundle adjustment software *Australis 6.0*. 27 images of our control point field from different viewpoints are used to achieve a reliable camera calibration (see Figure 2). It can be assumed that the image axes are perfectly orthogonal and the image aspect ratio is 1.0. The radial distortion with a maximum displacement of 1.8 pixels is eliminated by resampling all images using bicubic interpolation.

Figure 2: Control point field for camera calibration and the *KUKA* robot arm with the mounted digital camera.

We prefer an algebraic representation of the orientation data instead of the physical model to simplify the transformation from sensor coordinates to image coordinates (Hartley and Zisserman, 2000). The imaging geometry is modeled in terms of line-preserving 3×4 projection matrices $\mathbf{P}$ and a constant calibration matrix $\mathbf{K}$ with a principle distance in pixels:

$$\mathbf{K} = \begin{bmatrix} 2909.1 & 0 & 748.4 \\ 0 & 2909.1 & 408.7 \\ 0 & 0 & 1 \end{bmatrix} \quad (1)$$

$\mathbf{P}$ is a composite matrix computed from rotation, translation and calibration matrices. The image coordinates (x_i, y_i) of a given 3D point (X, Y, Z) is computed as follows.

$$w_i \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = \mathbf{P} \times \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \quad (2)$$

2.2.2 Texture Projector A *Sharp XG-P10XE* LCD-Projector with a resolution of 1024×768 pixel is used to map various textures onto the object (see Figure 3). Calibration of the LCD-projector is also necessary for some of the surface reconstruction algorithms such as shape-from-texture (see Section 3.3). The interior orientation is obtained from the technical reference sheet of the LCD-projector and transformed into pixel metric. The exterior orientation, however, must be computed manually. Again, we used the *Australis* software to compute the exterior orientation by modeling the LCD-projector as a special camera. The estimated orientation is verified by reprojecting a synthetic chessboard structure for which the 3D coordinates are already known.

Figure 3: *Sharp XG-P10XE* LCD-projector and *Liesegang 3600AV* slide-projector.

2.2.3 Light Projector The LCD-projector produces raster artifacts with blank texture images. So, an additional slide-projector is required to obtain images suitable for photogrammetric stereo. In our setup we use a *Liesegang 3600AV* model slide-projector to model a point light source (see Figure 4). Slide-projector positions are measured manually and verified by analyzing the highlights on five spheres, which are placed in the scene for the point cloud registration of the laser scanner. (see Figure 4)

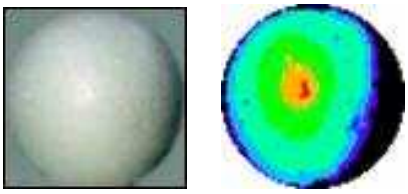


Figure 4: Styrofoam sphere with the estimated highlight position to verify the light direction.

2.3 Acquisition

The digital camera is firmly mounted on a *KUKA KR C1* industrial robot arm (see Figure 2). Due to various safety reasons, we have adjusted the camera and captured the images remotely through a firewire connection between a notebook and the camera. This also prevents the possible vibrations during image acquisition and improves the image quality. The robot enables a

stable and repeatable camera motion with a relative precision of 0.1 mm. Images are captured at 13 different positions, which have a horizontal baseline of approximately 1 m and a vertical baseline of 0.5 m. (see Figure 5).

Figure 5: Orientation of the reference image (red), narrow- and wide-baseline stereo images (green) and the linear motion sequence (yellow).

A computer-controlled *MICOS Pollux* high resolution positioning device (turntable) is configured to simulate a camera motion around the scene. Rotating the turntable in 9-degrees-steps, we acquired 40 images of the object (see Figure 6). It was not possible to rotate the LCD-projector and the slide-projector synchronous with the turntable, so we used the ambient light in the room and projected no texture on the object for the camera ring images.

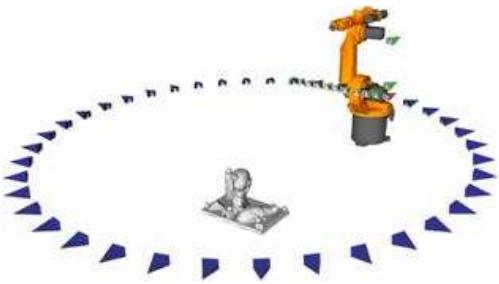


Figure 6: Overview of the imaging geometry. The ring (blue) is simulated using a computer controlled turntable.

2.4 Ground Truth

A *Mensi S25* short range triangulation laser scanner (Böhler and Marbs, 2002) is used to generate the independent ground truth information (see Figure 7). The accuracy of the scanner is about 0.8 mm vertical and 0.2 mm horizontal at a distance of 4 m orthogonal to range. The depth accuracy at this distance (4 m) is about 0.4 mm.

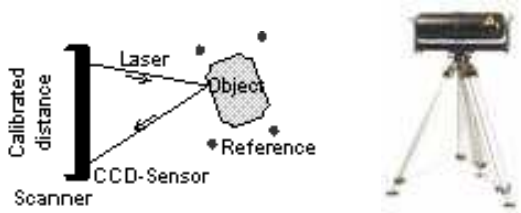


Figure 7: Triangulation principle of *Mensi S25* laser scanner.

The ground truth is acquired from 7 viewing directions in order to generate a dense 3D point cloud for the entire scene. Having registered the individual views, background and noisy points are manually removed. The registered and segmented 3D point cloud has a resolution of 2-3 mm (see Figure 8). Based on the scanned

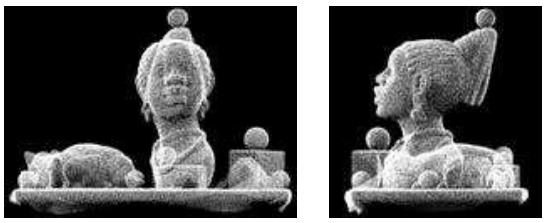
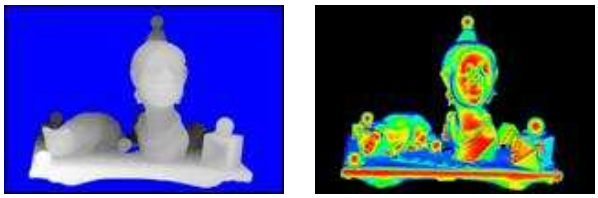


Figure 8: Laser scanner point cloud of the ground truth.

3D data, we derived two ground truth data maps for the reference image: a normalized depth map, where the gray value of each pixel is related to a depth value of the scene, and a color-coded surface orientation map, where the color of each pixel is related to an orientation vector (see Figure 9).



(a) Normalized depth map (b) Surface orientation map

Figure 9: Ground truth data for the reference image.

3 SHAPE-FROM-X ALGORITHMS

The basic idea of all image-based surface reconstruction techniques is that the color of light coming from a particular point in a scene can be rewritten as a function of the environment parameters, such as settings and the positioning of camera, lighting conditions, reflectance properties of surface and background, etc. In the following sections, we explain briefly the investigated surface reconstruction methods.

3.1 Shape-from-Stereo

Stereo is a well-known technique and works analogue to human vision. The scene must be observed from different viewpoints and for corresponding homologous image points the 3D object point is computed using the orientation information. This process is called triangulation. The correspondence problem is the most challenging part of this algorithm, especially for untextured and occluded areas. Various global optimization strategies, e.g. dynamic programming (Lei et al., n.d.), graph cuts (Kolmogorov and Zabih, 2001), belief propagation (Klaus et al., 2006) or semi-global matching (Hirschmüller, 2006), are used to achieve state-of-the-art results. Standardized test data for binocular stereo is already available (Scharstein and Szeliski, 2002, Courtney et al., 1996) including a comfortable online evaluation of the results.

3.1.1 Trinocular Stereo A binocular image match can be verified using a third image. Given two corresponding points the position in the third image can be predicted. The similarity of this point may support or reject the match. The convergent stereo triplets of our benchmark dataset are rectified, so that they correspond to the stereo normal case (Heinrichs and Rodehorst, 2006). The important advantage of this method is that the correspondence analysis is simplified to a linear search along one of the image axes (see Figure 10-11). An additional property is that the differences (disparities) between horizontal and vertical corresponding points are equal.

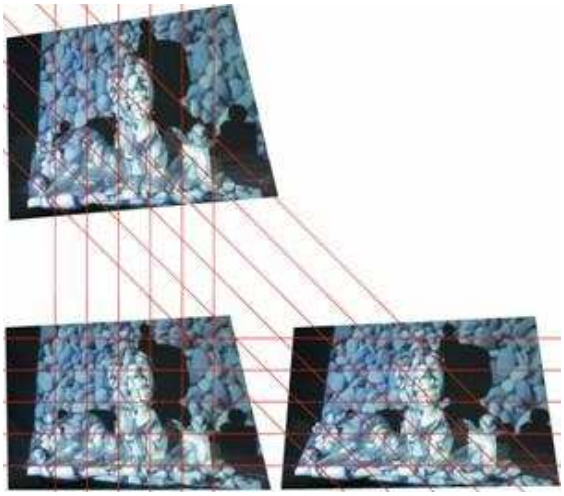


Figure 10: Example of a rectified narrow-baseline triplet with overlaid epipolar lines (23 textured image triplets)

3.1.2 Narrow-baseline Stereo A smaller distance between the cameras leads to very similar images and simplifies the correspondence analysis. However, as the triangulation angle gets smaller, accurate spatial intersection gets harder. This phenomenon is also called glancing intersection (see Figure 10).

3.1.3 Wide-baseline Stereo A larger distance between the cameras makes the spatial intersection easier. But, in this case, perspective distortions and stronger occlusions make the correspondence analysis more difficult (see Figure 11).

3.1.4 Multi-view Stereo (Shape-from-Motion) Multi-view stereo combines the advantages of narrow- and wide-baseline stereo. Using neighboring images, the correspondence analysis is simplified, and the baseline for a spatial intersection is extended by using feature tracking. In (Seitz et al., 2006) standardized test data for multi-view stereo with an online evaluation was made available.

3.2 Shape-from-Photoconsistency

3.2.1 Shape-from-Silhouette (Space Carving) Shape-from-silhouette is a surface reconstruction method which constructs a 3D shape estimate of an object using multiple contour images of the object. The output is known as the visual hull. This technique still suffers from concavities and insufficient views. A simple approach is the volumetric reconstruction using voxel space carving (Martin and Aggarwal, 1983). The final reconstruction is still coarse. A more accurate representation is possible with marching intersections (Tarini et al., 2002) or polygonal reconstructions using generalized cone intersections (Matusik et al., 2000).

3.2.2 Shape-from-Photoconsistency (Voxel Coloring) The analysis of consistent scene colors and texture can be used to refine the visual hull (Seitz and Dyer, 1999). Assuming Lambertian reflection and textured surfaces a reconstruction of concave areas is possible.

3.2.3 Shape-from-Shadow (Shadow Carving) This approach analyzes self-shadows on the surface which may indicate concavities (Savarese et al., n.d.). Here, the illumination direction has to be given, and the detection as well as the categorization of shadows is difficult.

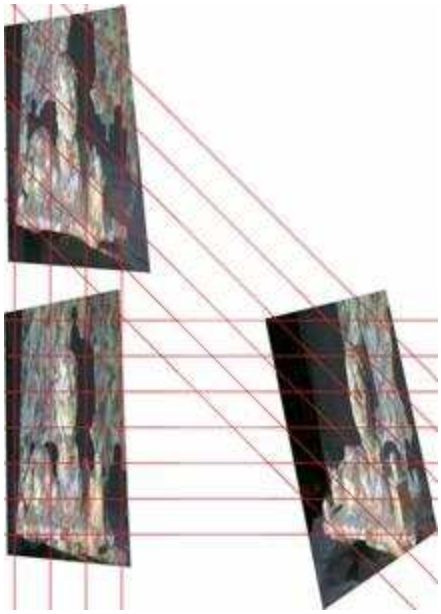


Figure 11: Example of a rectified wide-baseline triplet (23 image textured triplets)

Figure 12: Examples of the image sequence acquired with a linear camera motion (253 textured images)

3.3 Shape-from-Texture

In shape-from-texture, known texture patterns are projected onto the scene, and surface of the scene is estimated by observing the distortions of these patterns in the acquired images. Structured light is one of the commonly used cases of this method (Scharstein and Szeliski, 2003). It is based on active triangulation by replacing the second stereo camera with a calibrated LCD-projector. Multiple stripes are projected onto the scene, in order to distinguish between stripes they are coded either with different brightness or different colors (see Figure 14-15). Therefore, correspondence problem marginally solved. The depth information can be computed out of the distortion along the detected profiles.

We projected a standard binary pattern and a rainbow like color spectrum onto the scene. Every stripe of length n in the color spectrum S_{RGB} is generated using the below equations, where i denotes the raster position in the spectrum image and G_{MAX} denotes the maximum intensity value in every color channel (Koschan and Rodehorst, 1997).

$$\begin{aligned} S_R &= sin(\tfrac{i}{n} \cdot \pi) \cdot (\tfrac{G_{MAX}}{2} - 1) + \tfrac{G_{MAX}}{2} \\ S_G &= sin((\tfrac{2}{3} + \tfrac{i}{n}) \cdot \pi) \cdot (\tfrac{G_{MAX}}{2} - 1) + \tfrac{G_{MAX}}{2}, i = \{0, .., n\} \\ S_B &= sin((\tfrac{4}{3} + \tfrac{i}{n}) \cdot \pi) \cdot (\tfrac{G_{MAX}}{2} - 1) + \tfrac{G_{MAX}}{2} \end{aligned} \quad (3)$$

3.4 Shape-from-Shading

Shape-from-shading (Horn and Brook, 1989) is a single image technique to estimate a 3D surface with Lambertian reflectance

Figure 13: Examples of the turntable sequence with segmented contour images (40 images)

Figure 14: Examples of binary coded textures (9 images with horizontal and 10 images with vertical stripes)

properties from a known illumination direction. Photometric stereo is an extension that uses at least three monoscopic images with different illumination directions to achieve a unique result (Klette et al., 1999). These reflectance based methods determine surface orientations instead of depth information.

3.5 Shape-from-(De-)Focus

The last approach uses monoscopic image sequences with varying focal length (Favaro and Soatto, 2005). A segmentation of sharp and blurred image parts provides information about the actual focused depth. This method requires special sensor properties with an extreme small depth of field (e.g. microscope).

4 EVALUATION AND RESULTS

The surface reconstruction methods, which are investigated in this study, have various output formats such as disparity maps, depth maps, volumetric data or surface orientation maps (see Table 2). In order to compare these different outcomes quantitatively, one should first convert them to a comparable form. Fortunately, except for the surface orientation map, all other formats can easily be converted to a depth map, if image orientation is available. As we already provide the image orientation, quantitative comparison can be performed through depth maps. So, for evaluating reflectance based methods the surface orientation maps are used, whereas for the other methods depth maps are preferred. Once the reconstruction outcome is converted into a depth/orientation map, it is possible to compare it with the ground truth depth/orientation map. This comparison is done according to two criteria. The first criterion is accuracy. Accuracy is a measure of closeness of the values of the reconstruction and the ground truth. Second criterion is completeness. It is a measure of overall success of the algorithm. The more the scene is modeled, the more complete is the reconstruction.

Method	Output
Stereo	Disparity map, depth map
Motion	Polyhedron, point cloud
Silhouette	Volumetric data, polyhedron
Texture	Disparity map, depth map
Shading	Surface orientation, needle map
Focus	Depth map

Table 2: Output of investigated methods.

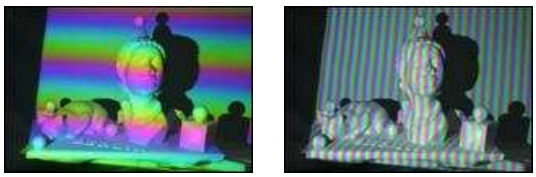


Figure 15: Examples of color spectrum textures (9 images with horizontal and 10 images with vertical pattern)

(a) Left (b) Top (c) Right

Figure 16: Reference image under different illumination configurations (3 Lambertian-like images)

Let us denote the ground truth map with G and the reconstructed map with R. The accuracy of the reconstruction is computed using the difference between these two maps as shown in Equation 4. In this equation M is the set of pixels used in evaluation, n is the total number of pixels in M, and g is the normalization factor. Normalization factor is either the maximum depth value (i.e. $g = 255$) or the maximum norm of the difference vector given in the equation (i. e. $g = 2$).

$$a_{R,M} = 1 - \frac{1}{n} \times \sum_{i,j \epsilon M} \left(\frac{|R(i,j) - G(i,j)|}{g} \right) \qquad (4)$$

The accuracy term can be used to rank different reconstruction results. However, accuracy alone would not be a sufficient for a fair ranking. If there are pixels on the map, whose depth/orientation values cannot be correctly reconstructed, this should also be taken into consideration. Completeness is defined as the ratio of well-reconstructed pixels to all pixels as given in Equation 5. δ is a threshold for error tolerance. The accuracy and completeness of an implementation of the space carving algorithm is given in Table 3 as an example. Reconstruction result of an implementation of the narrow-baseline stereo algorithm is illustrated in Figure 18

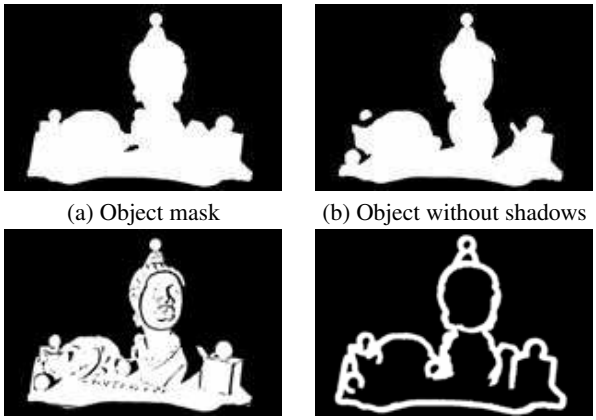
$$c_{R,M} = \frac{1}{n} \times \sum_{i,j \epsilon M} \left(\frac{|R(i,j) - G(i,j)|}{g} < \delta \right) \qquad (5)$$

Using the formulas provided above it is possible to measure the accuracy and the completeness of a reconstruction at specific regions. This allows us to investigate the success of each method in reconstructing challenging cases, such as self occlusion, concavities, sharp surface discontinuities etc. Thus, mostly inspired by (Courtney et al., 1996), we provide masks (see Figure 17), which can be used to compute the accuracy and completeness according to the following criteria.

Uniformity: Regions where the deviation of gradients of the pixels is less than a given threshold.

Occlusion: Regions that are occluded in the matching image.

Depth discontinuity: Regions where the deviation of depth values of the pixels is less than a given threshold.



(a) Object mask (b) Object without shadows

(c) Uniformity mask (d) Depth discontinuity mask

Figure 17: Masks for investigating detailed aspects.

Mask	Accuracy	Completeness $\delta = 0.10$
Object	0.9787	0.6998
Object without shadows	0.9883	0.7255
Uniformity mask	0.9864	0.7272
Depth discontinuity	0.9185	0.6557

Table 3: Accuracy and completeness of an implementation of the space carving algorithm

5 CONCLUSIONS AND FUTURE WORK

In this study, we introduce a true benchmark dataset for performance evaluation of shape-from-X algorithms and a test procedure for evaluating the reconstruction results. Our aim in this work is to support an objective comparison of different approaches and to provide an informative basis for the combination of reconstruction methods. Researchers are invited to download this benchmark dataset data from our web server and return their results to see how successful or unsuccessful their method with respect to other methods is.

REFERENCES

Böhler, W. and Marbs, A., 2002. 3d scanning instruments. In: International Workshop on Scanning for Cultural Heritage Recording, Corfu, Greece.

Courtney, P., Thacker, N. and Clark, A., 1996. Algorithmic modelling for performance evaluation. In: European Conference on Computer Vision Workshop on Performance Characteristics of Vision Algorithms, Cambridge.

Favaro, P. and Soatto, S., 2005. A geometric approach to shape from defocus. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Foerstner, W., 1996. 10 pros and cons against performance characterisation of vision algorithms. In: European Conference on Computer Vision Workshop on Performance Characteristics of Vision Algorithms, Cambridge.

Hartley, R. I. and Zisserman, A., 2000. Multiple view geometry in computer vision.

Heinrichs, M. and Rodehorst, V., 2006. Trinocular rectification for various camera setups. In: Symposium of ISPRS Commission III - Photogrammetric Computer Vision, Bonn, pp. 43–48.

Hirschmüller, H., 2006. Stereo vision in structured environments by consistent semi-global matching. In: IEEE Conference on Computer Vision and Pattern Recognition, New York, pp. 2386–2393.

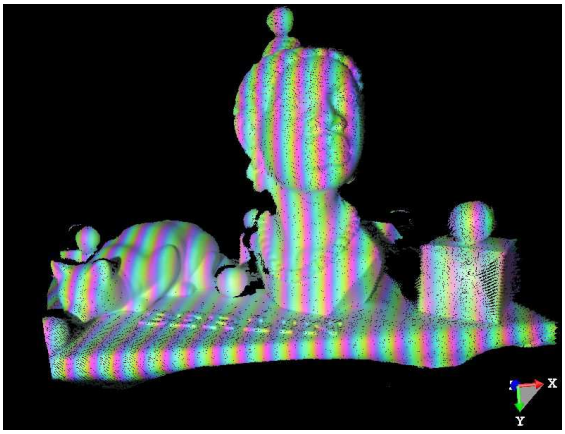


Figure 18: Reconstruction result of an implementation of the narrow-baseline stereo algorithm

Horn, B. K. P. and Brook, M. J., 1989. Shape from shading, mit press, cambridge, ma.

Klaus, A., Sormann, M. and Karner, K., 2006. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In: International Conference on Pattern Recognition, Hong Kong, pp. 15–18.

Klette, R., Kozera, R. and Schlüns, K., 1999. Reflectance-based shape recovery, handbook of computer vision, vol. 2, academic press, 531-590.

Kolmogorov, V. and Zabih, R., 2001. Computing visual correspondence with occlusions via graph cuts. In: International Conference on Computer Vision, pp. 508–515.

Koschan, A. and Rodehorst, V., 1997. Dense depth maps by active color illumination and image pyramids. In: In: F. Solina, W.G. Kropatsch, R. Klette, R. Bajcsy (Eds.), Advances in Computer Vision, Springer-Verlag,, Vienna, pp. 137–148.

Lei, C., Selzer, J. and Yang, Y. H., n.d. Region-tree based stereo using dynamic programming optimization, ieee conf. on computer vision and pattern recognition cvpr'06, new york, ny, usa, pp. 2378-2385, 2006.

Martin, W. N. and Aggarwal, J. K., 1983. Volumetric description of objects from multiple views. IEEE Transactions on Pattern Analysis and Machine Intelligence 5(2), pp. 150–158.

Matusik, W., Buehler, C., Raskar, R., Gortler, S. J. and McMillan, L., 2000. Image based visual hulls. In: ACM Computer Graphics SIGGRAPH, pp. 368–374.

Savarese, S., Rushmeier, H., Bernardini, F. and Perona, P., n.d. Shadow carving, proc. of the ninth ieee interational conference on computer vision, vancouver, ca, 2001.

Scharstein, D. and Szeliski, R., 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision 47((1/2/3)), pp. 7–42.

Scharstein, D. and Szeliski, R., 2003. High-accuracy stereo depth maps using structured light. In: IEEE Conference on Computer Vision and Pattern Recognition, Madison, pp. 195–202.

Seitz, S. and Dyer, C., 1999. Photorealistic scene reconstruction by voxel coloring. International Journal of Computer Vision 35(2), pp. 151–173.

Seitz, S. M., Curless, B., Diebel, J., Scharstein, D. and Szeliski, R., 2006. A comparison and evaluation of multi-view stereo reconstruction algorithms. In: IEEE Conference on Computer Vision and Pattern Recognition, New York, pp. 519–526.

Tarini, M., Callieri, M., Montani, C., Rocchini, C., Olsson, K. and Persson, T., 2002. Marching intersections: An efficient approach to shape-from-silhouette. In: IEEE Workshop on Vision, Modeling, and Visualization, Erlangen, pp. 283–290.